Optimización de las técnicas utilizadas para el proceso de descongelación de pesca de Jurel mediante modelos predictivos de datos

Andreina Barrera
Luis Morán
Facultad de Posgrado UEES
Maestría en Inteligencia de Negocios y Ciencia de Datos
23 de noviembre del 2024
Palabras clave: Predicción, árboles de decisión, KNN, optimización procesos, jurel

Tabla de contenido

A. Resumen Ejecutivo 1
B. Resumen Ejecutivo Gráfico 1
C. Detalle Técnico 2
1. Fuente de Datos 2
2. Análisis exploratorio 3
3. Técnica 4
D. Resultados 5
E. Referencias bibliográficas 8
F. Anexos 9

Tabla de Ilustraciones

Figura 1: Medidas de rendimiento del modelo con 2 técnicas 1
Figura 2: Variables relevantes del modelo 2
Figura 3: Balance de datos de la variable dependiente 3
Figura 4: Casos de pérdida y no pérdida con relación a 1% 3
Figura 5: Promedio de Porcentaje de pérdida 3
Figura 6: Casos de porcentaje de pérdida menor a 1% por mes 4
Figura 7: Diagrama de caja Temperatura Ambiente 4
Figura 8: Árbol de decisión caso de estudio 5
Figura 9: Curvas ROC de ambos modelos 6
Tabla 1: Resumen de medidas de rendimiento 6

A. Resumen Ejecutivo

De acuerdo con el informe del 2024, El estado mundial de la pesca y la acuicultura (SOFIA), en los países en desarrollo, un porcentaje significativo de desperdicio se genera en la etapa posterior a la cosecha, especialmente debido a limitaciones en la infraestructura de procesamiento y almacenamiento donde se desaprovechan grandes recursos; por ello, las compañías deben optimizar su proceso productivo. En Chile, la empresa Pesquera S.A. ha identificado una merma significativa en su proceso de descongelamiento de pesca del Scomber Scrombrus y el Jurel (Trachurus Murphy). Con el objetivo de reducir el porcentaje de pérdidas en merma al 1%, la empresa ha llevado a cabo un análisis detallado recopilando datos clave sobre las variables que influyen en el descongelamiento de sus productos entre enero y agosto de 2024. Este enfoque busca implementar mejoras para optimizar su operación y mejorar la eficiencia.

El objetivo principal del presente proyecto es crear un modelo de predicción para determinar si en un proceso de descongelamiento se cumplirá o no con el porcentaje de pérdida establecido. Para ello, se usará la técnica de árboles de decisión, debido a su fácil interpretación y a su capacidad para determinar cuáles son las variables más importantes para la predicción.

Entre los resultados más importantes destaca que la variable más importante para la predicción fue la temperatura ambiente, seguida del peso promedio. El modelo de predicción resultó tener un porcentaje de precisión del 86% al predecir si un proceso cumplirá con la meta del porcentaje de pérdida.

B. Resumen Ejecutivo Gráfico

Durante la ejecución de las dos técnicas entre árboles de decisión y KNN se obtuvo que los datos y el modelo poseen buenos parámetros de rendimiento, aunque el árbol de decisión se destaca como una opción ligeramente mejor debido a que se obtiene un error del 14% en las predicciones (KNN: 15%), así como también se obtiene una precisión en la clasificación correcta de un 86% de los casos (KNN: 85%). Finalmente, el F1-Score de un 86% indica un buen equilibrio en el porcentaje de predicciones positivas son correctas y el porcentaje de las instancias positivas reales que son correctamente identificadas. (KNN 84%)

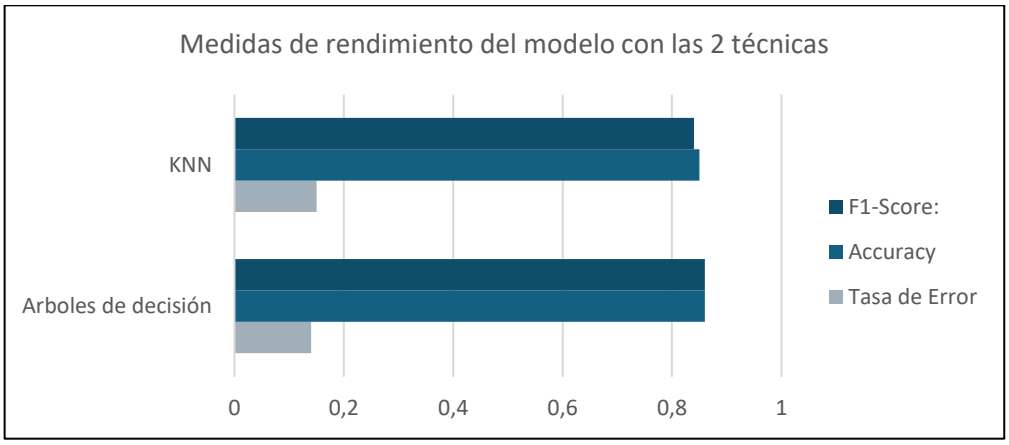


Figura 1: Medidas de rendimiento del modelo con 2 técnicas

Para determinar si el árbol tiene capacidad de predecir los resultados de nuevos procesos de descongelamiento, se dividió la base de datos en entrenamiento (70%) y prueba (30%). Una vez obtenido, el árbol indica que las variables relevantes son:



Figura 2: Variables relevantes del modelo

Considerando las variables obtenidas, el modelo se encuentra en concordancia con los factores químicos asociados al procesamiento y producción de productos a base de materia prima de peces. Según Otero (2014), entre los criterios de evaluación de la calidad de la materia prima se encuentra la temperatura en la que se realiza el procesado y, en el caso de estudio, se observó que durante los meses de junio y julio se logró una notable reducción en el porcentaje de merma durante el periodo de descongelamiento en la producción pesquera. Este resultado se atribuye, en parte, a la disminución de la temperatura ambiente en las costas chilenas, que alcanzó un promedio de 84.2 grados Fahrenheit. La disminución en la temperatura ambiental favoreció a un proceso de descongelamiento más controlado, alineándose con el objetivo estratégico de minimizar las pérdidas y maximizar el aprovechamiento del producto.

C. Detalle Técnico

1. Fuente de Datos

Los datos corresponden a los recopilados por la empresa Pesquera S.A. durante sus procesos de descongelamiento, entre los meses de enero – agosto del 2024. En ellos se incluyen datos de las variables que intervienen durante el proceso. La información de cada una de las variables se encuentra en el Anexo 1.

Considerando la limpieza de datos como parte esencial de los procesos de integración y modelado de datos en almacenes de datos, según Kimball & Ross (2002), la base de datos cuenta con registros de 932 procesos de descongelado durante el periodo de tiempo mencionado previamente del cual se realizó una revisión exhaustiva e integral de la totalidad de los datos. Durante este proceso, se verificó que no existen registros duplicados ni datos vacíos que puedan comprometer la integridad del análisis. Así mismo, se realizó la revisión sobre el balance de los datos para identificar si es necesario para balancear conjuntos de datos como lo indica Chawla et al. (2002). Como resultado del procedimiento en mención no se identificaron valores atípicos ni inconsistencias que pudieran considerarse inusuales asegurando un adecuado balance en todas las variables analizadas.

Como se puede observar en el gráfico a continuación, se obtuvo una relación de 42-57, lo cual se considera dentro del estándar del balanceo 60-40.

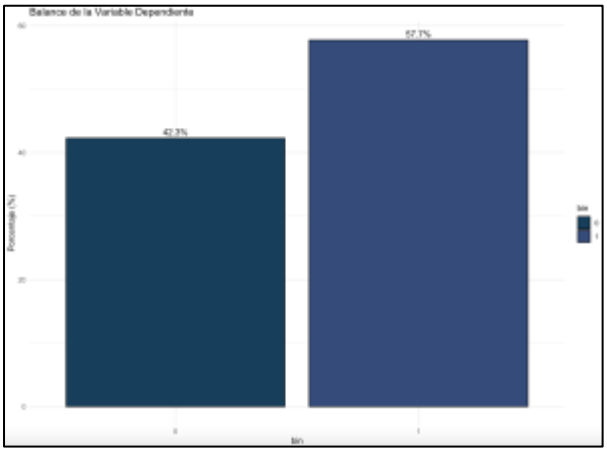


Figura 3: Balance de datos de la variable dependiente

Lo descrito anteriormente permite concluir que la base de datos cumple con los estándares requeridos para su correcta utilización en cualquier proceso analítico, garantizando resultados confiables y representativos.

2. Análisis exploratorio

Actualmente la empresa obtiene un promedio de porcentaje de pérdida de un 1.16%, a pesar de que el 57.01% de sus procesos resultaron en pérdidas bajo el estándar establecido. Esto quiere decir que los porcentajes de pérdida en aquellos casos donde se supera el umbral del 1% son significativamente altos.



Figura 4: Casos de pérdida y no pérdida con relación a 1%



Figura 5: Promedio de Porcentaje de pérdida

Analizando la base de datos según los meses, se detecta que la mayor parte de procesos con porcentajes de merma menores al 1% sucedieron en los meses de junio y julio, durante los cuales se produjo una disminución en la temperatura ambiente en las costas chilenas. Esto puede indicar que un aumento en dicha temperatura puede afectar los procesos de descongelamiento y al porcentaje de merma.



Figura 6: Casos de porcentaje de pérdida menor a 1% por mes

Durante el análisis exploratorio de datos, se puede observar cómo al relacionar la variable binaria dependiente con la temperatura ambiente, hay una clara división entre los casos de éxito y de fracaso. Cuando la temperatura ambiente es alta, la merma tiende a ser mayor a 1%; sin embargo, cuando la temperatura es baja, la merma tiende a ser menor a 1%. Este resultado va acorde a lo encontrado en el árbol de decisión y refuerza la importancia de esa variable.

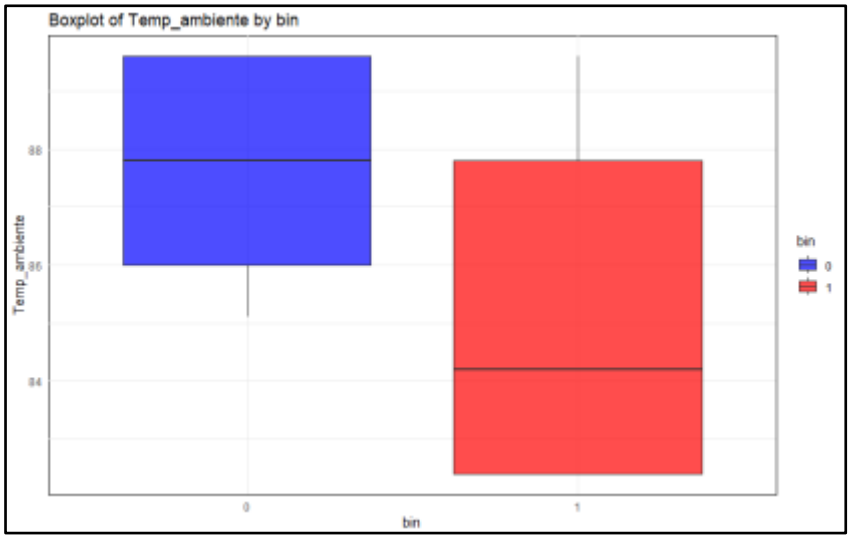


Figura 7: Diagrama de caja Temperatura Ambiente

3. Técnica

Para generar los modelos predictivos se utilizarán las técnicas de árboles de decisión y KNN (K-Nearest Neighbors) de clasificación, ya que la variable dependiente es una binaria. Se han elegido estas técnicas ya que el árbol de decisión muestra las predicciones de una manera gráfica y sencilla de interpretar y son herramientas fáciles de interpretar porque reflejan explícitamente las reglas que conducen a una clasificación o decisión, lo cual es importante para que todos quienes forman parte

del proceso puedan utilizarla, mientras que el uso de otra técnica de predicción de clasificación como KNN nos permitirá comparar resultados y medidas de rendimiento para elegir el mejor modelo predictivo. Para la aplicación de estas técnicas se utilizó el software estadístico R.

Se verificó el balanceo de las clases en la variable dependiente ya que los árboles de decisión pueden ser sensibles a datos desbalanceados. Se comprobó que los datos tenían un buen balance, sin embargo, se decidió utilizar undersampling para que el balance sea 50/50 y que los modelos tengan igual representación para ambas clases. Para el caso del modelo KNN, se realizó además la normalización de las variables mediante Z-score, ya que existían muchas variaciones de escala que podrían afectar al modelo. Gracias a la buena cantidad de datos, se pudo dividir la base en entrenamiento (70%) y prueba (30%), lo cual nos permitirá usar los datos tanto para crear ambos modelos como para verificar si pueden ser usados para predecir resultados futuros.

Para optimizar los modelos, se utilizó validación cruzada para encontrar el k óptimo en el caso de KNN y el Parámetro de Complejidad óptimo en el caso del árbol de decisión. El árbol de decisión fue podado teniendo en cuenta este parámetro. La optimización se realiza para evitar problemas de sobreajuste a los modelos y tengan una buena capacidad de generalización. Una vez optimizado cada modelo, se realizan predicciones cuyos resultados se evalúan con los datos de prueba y se crea una matriz de confusión para determinar los niveles de precisión y error. Estos valores se comparan para determinar cuál de las dos técnicas genera el mejor modelo de predicción para este caso.

D. Resultados

Una vez obtenidos el modelo de árbol de decisión y de KNN, se realizó la optimización de estos mediante la técnica de Validación Cruzada. Los resultados de cada proceso se encuentran en el Anexo 2 y Anexo 3.

Gracias a estos resultados podemos determinar que el tamaño óptimo del árbol son 8 nodos, mientras que el k óptimo para KNN es 20. Teniendo esto en cuenta, el árbol de decisión obtenido es el siguiente:

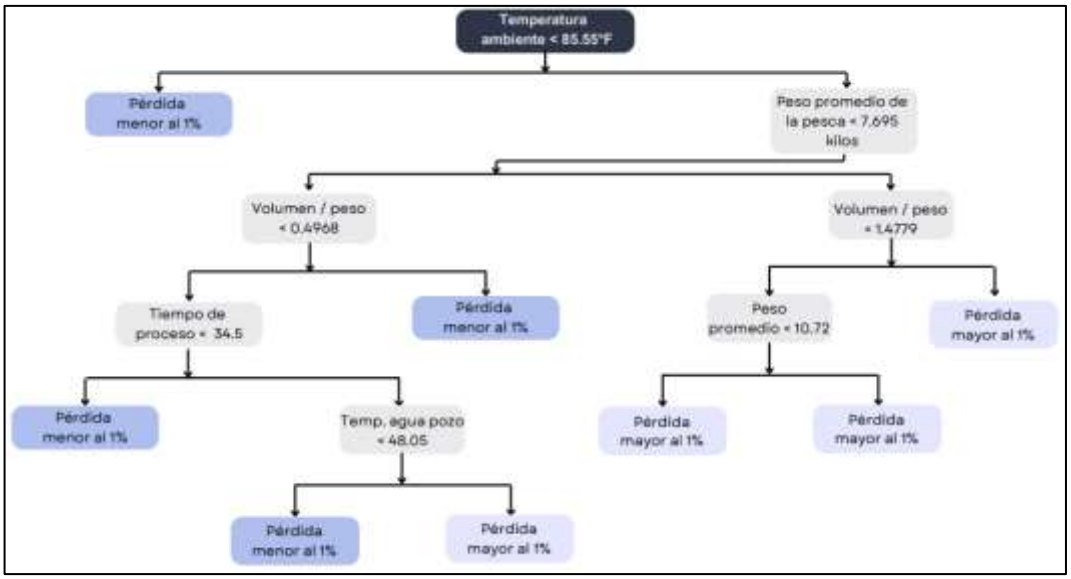


Figura 8: Árbol de decisión caso de estudio

El árbol de decisión nos permite determinar que las variables más importantes son la Temperatura Ambiente, seguida del Peso Promedio, Volumen por tonelada, Tiempo del Proceso y Temperatura de agua del pozo.

Una vez obtenidos los modelos mediante el uso de los parámetros óptimos y los datos de entrenamiento, se realizan predicciones y se comparan con los datos de prueba para medir el rendimiento de los modelos y elegir el mejor. Según Han et al. (2012), se utilizan las medidas de rendimiento como la precisión, el recall, la F1-Score y la curva ROC, en el contexto del análisis predictivo para evaluar la confiabilidad del modelo. Las medidas de rendimiento de los modelos se obtuvieron de una matriz de confusión con los siguientes resultados:

Medida de rendimiento	Árbol de decisión	KNN
Tasa de Error	0,14	0,15
Accuracy	0,86	0,85
F1-Score:	0,86	0,84
Valor AUC	0,89	0,92

Tabla 1: Resumen de medidas de rendimiento

En general, ambos modelos obtuvieron medidas de rendimiento muy similares, aunque el árbol de decisión arrojó valores ligeramente mejores. La tasa de error de ambos modelos es baja, mientras que el F1-Score nos indica que los modelos muestran un buen balance entre ambas clases de la variable dependiente. Otro valor importante es el Valor AUC, el cual evalúa la bondad de los modelos. Para ambos casos, el valor AUC es considerado alto, ya que se encuentra cerca de 1. Un modelo con un valor AUC alto indica que tiene un gran rendimiento y gran capacidad para distinguir entre las clases. Para el caso de KNN, este valor resultó un poco por encima del obtenido del árbol de decisión. En cuanto a las curvas ROC para cada modelo, ambas se encuentran muy alejadas de la diagonal principal, lo cual se considera que corresponde a modelos de gran rendimiento. La curva ROC del modelo KNN se encuentra un poco por encima que la del árbol de decisión, debido a su valor AUC ligeramente superior.

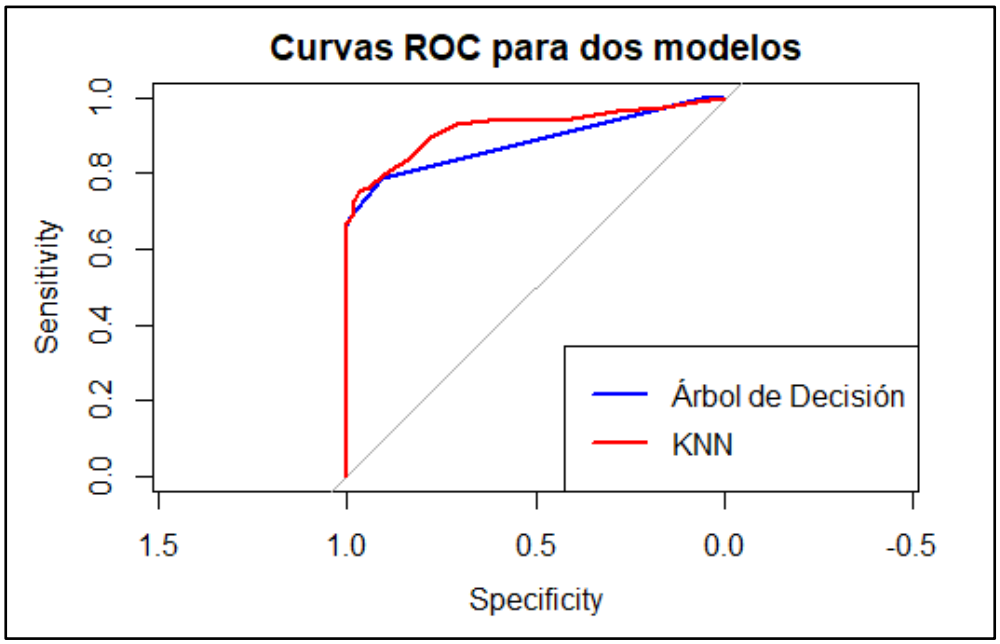


Figura 9: Curvas ROC de ambos modelos

Como conclusión, se decide utilizar el modelo de árbol de decisión para predecir si se obtendrá el porcentaje de pérdida deseado para este caso. Aunque las medidas de rendimiento son muy similares, el árbol de decisión ofrece la ventaja de poder realizar las predicciones mediante el gráfico establecido, sin la necesidad de utilizar un software. Además, es un modelo simple y fácil de interpretar tanto para los directivos de la empresa como para los trabajadores en general.

E. Referencias bibliográficas

FAO Knowledge Repository. (2024). Fao.org. https://openknowledge.fao.org/items/06690fd0-d133-424c-9673-1849e414543d

Otero, M. (2014). Informe Descriptivo Del Procesamiento De Lomos De Atun Precocidos Congelados Sellados Al Vacío. Universidad Nacional De Piura. Recuperado de: https://core.ac.uk/download/pdf/250077163.pdf

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.* *Wiley.* http://160592857366.free.fr/joe/ebooks/ShareData/The%20Data%20Warehouse%20Toolkit.pdf

Universidad Espíritu Santo. (s. f.). Manual de Normas APA para estudiantes 7ma edición. En *Online*.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (Third Edition). https://www.sciencedirect.com/book/9780123748560/data-mining-practical-machine-learning-tools-and-techniques#book-info

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

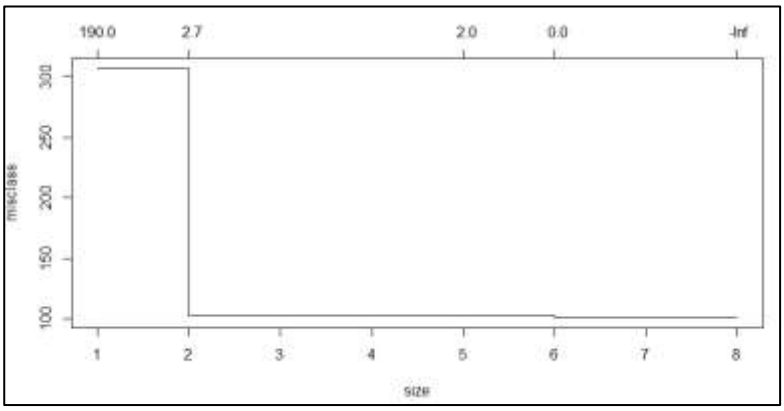
Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques: A volume in The Morgan Kaufmann Series in Data Management Systems* (Third Edition). https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

F. Anexos

Anexo 1. Operacionalización de las variables

Variable	Descripción	Tipo de Variable	Medida
Cámara	Pozo en el que se realizó el descongelamiento	Categórica	
Peso_Prom	Peso promedio del marisco	Numérica	kilos
Presion_aire	Presión del aire	Numérica	psi
Temp_agua_pozo	Temperatura del agua en el descongelamiento	Numérica	°F
T_Proceso	Tiempo del proceso de descongelado	Numérica	min
Vol_peso	Volumen de marisco por tonelada	Numérica	M3/mt
Indice_Rotacion	Revoluciones por hora	Numérica	R/h
Temp_final	Temperatura final del proceso	Numérica	°F
Temp_ambiente	Temperatura ambiente durante el proceso	Numérica	°F
Bin (Porcentaje de pérdida)	0 si el porcentaje de pérdida fue mayor a 1%, 1 si fue menor a 1%.	Binaria	

Anexo 2. Curva de Rendimiento de Árboles de Decisión con Validación Cruzada



Anexo 3. Métricas de Rendimiento de valores k con Validación Cruzada

k	Accuracy	Kappa
2	0.7646982	0.5291848
3	0.8240825	0.6480630
4	0.8151551	0.6303457
5	0.8206097	0.6411411
18	0.8257672	0.6512994
19	0.8295707	0.6589030
20	**0.8349603**	**0.6697098**
21	0.8331085	0.6660152
22	0.8313228	0.6623371
28	0.8276876	0.6550968
29	0.8295382	0.6587667
30	0.8312903	0.6622937